

## Robust training on approximated minimal-entropy set

Tianpei Xie, Nasser M. Nasrabadi, *Fellow, IEEE*, and Alfred O. Hero, *Fellow, IEEE*

**Abstract**—In this paper, we propose a general framework to learn a robust large-margin binary classifier when corrupt measurements, called anomalies, caused by sensor failure might be present in the training set. The goal is to minimize the generalization error of the classifier on non-corrupted measurements while controlling the false alarm rate associated with anomalous samples. By incorporating a non-parametric regularizer based on an empirical entropy estimator, we propose a Geometric-Entropy-Minimization regularized Maximum Entropy Discrimination (GEM-MED) method to learn to classify and detect anomalies in a joint manner. We demonstrate using simulated data and a real multimodal data set. Our GEM-MED method can yield improved performance over previous robust classification methods in terms of both classification accuracy and anomaly detection rate.

**Index Terms**—sensor failure, robust large-margin training, anomaly detection, maximum entropy discrimination.

## I. INTRODUCTION

Large margin classifiers, such as the support vector machine (SVM) [1] and the maximum entropy discrimination (MED) classifier [2], have enjoyed great popularity in the signal processing and machine learning communities due to their broad applicability, robust performance, and the availability of fast software implementations. When the training data is representative of the test data, the performance of MED/SVM has theoretical guarantees that have been validated in practice [1], [3], [4]. Moreover, since the decision boundary of the MED/SVM is solely defined by a few support vectors, the algorithm can tolerate random feature distortions and perturbations.

However, in many real applications, anomalous measurements are inherent to the data set due to strong environmental noise or possible sensor failures. Such anomalies arise in industrial process monitoring, video surveillance, tactical multi-modal sensing, robust spectrum sensing [5], [6], and, more generally, any application that involves unattended sensors in difficult environments (Fig. 1). Anomalous measurements

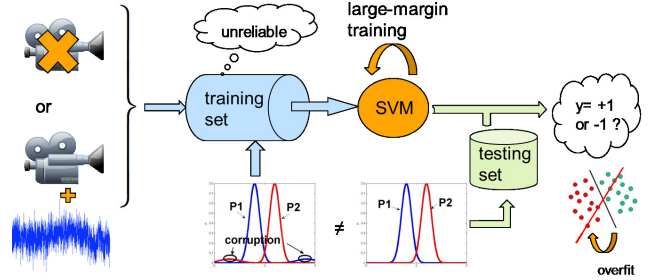


Fig. 1. Due to corruption in the training data the training and testing sample distributions are different from each other, which introduces errors into the decision boundary.

are understood to be observations that have been corrupted, incorrectly measured, mis-recorded, drawn from different environments than those intended, or occurring too rarely to be useful in training a classifier [7]. If not robustified to anomalous measurements, classification algorithms may suffer from severe degradation of performance. Therefore, when anomalous samples are likely, it is crucial to incorporate outlier detection into the classifier design. This paper provides a new robust approach to design outlier resistant large margin classifiers.

### A. Problem setting and our contributions

We divide the class of supervised training methods into four categories, according to how anomalies enter into different learning stages.

TABLE I  
CATEGORIES FOR SUPERVISED TRAINING ALGORITHMS VIA DIFFERENT  
ASSUMPTION OF ANOMALIES

	Training set (uncorrupted)	Training set (corrupted)
Test set (uncorrupted)	classical learning algorithms (e.g. [2], [8], [9])	Robust classification & training (e.g. [3], [4], [10]–[19], <b>this paper</b> )
Test set (corrupted)	anomaly detection (e.g. [20]–[23])	Domain adaptation & transfer learning (e.g. [24]–[26])

As shown in Table I, a majority of learning algorithms assume that the training and test samples follow the same nominal distribution and neither are corrupted by anomalies. Under this assumption, an empirical error minimization algorithm can achieve consistent performance on the test set. In the case that anomalies exist only in the test data, one can apply anomaly detection algorithms, e.g. [21]–[23], [27], to separate the anomalous samples from nominal ones. Under additional assumptions on the nominal set, these algorithms can effectively identify an anomalous sample under given false alarm rate and miss rate. Furthermore, in the case that both training

Tianpei Xie is with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, 48109, USA e-mail: (tianpei@umich.edu).

Nasser M. Nasrabadi is with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, US. email: nasser.nasrabadi@mail.wvu.edu

Alfred O. Hero is with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI., 48109, USA e-mail: (hero@umich.edu).

This research was supported by US Army Research Office (ARO) grants W11NF-11-1-103A1.

and test set are corrupted, possibly with different corruption rate, domain adaptation or transfer learning methods may be applied [24], [26], [28].

This paper falls into the category of *robust classification & training* in which possibly anomalous samples occur in the training set. Such a problem is relevant, for example, when high quality clean training data is too expensive or too difficult to obtain. In [3], [4], [12], the test set is assumed to be uncorrupted so that the test error provides an unbiased estimate of the generalization error on the *nominal data set*, which is a standard measure of performance for robust classifiers. We adopt this assumption, although we also evaluate the proposed robust classifier when the test set is also corrupted with limited corruption rate. Our *goal* is to train a classifier that minimizes the generalization error with respect to the nominal data distribution when the *training* set may be corrupted.

The area of robust classification has been thoroughly investigated in both theory [3], [4], [10], [12]–[14], [16] and applications [15], [17]–[19]. Tractable robust classifiers that identify and remove outliers, called the Ramp-Loss based learning methods, have been studied in [3], [11], [13], [15]. Among these methods, Xu et al. [13] proposed the Robust-Outlier-Detection (ROD) method as an outlier detection and removal algorithm using the soft margin framework. Training the ROD algorithm involves solving an optimization problem, for which dual solution is obtained via semi-definite programming (SDP). Like all the Ramp-Loss based learning models, this optimization is non-convex requiring random restarts to ensure a globally optimal solution [7], [17]. In this paper, in contrast to the models above, a *convex* framework for robust classification is proposed and a tractable algorithm is presented that finds the unique optimal solution of a penalized entropy-based objective function.

Our proposed algorithm is motivated by the basic principle underlying the so-called *minimal volume (MV) /minimal entropy (ME) set anomaly detection method* [20]–[23]. Such methods are expressly designed to detect anomalies in order to attain the lowest possible false alarm and miss probabilities. In machine learning, nonparametric algorithms are often preferred since they make fewer assumptions on the underlying distribution. Among these methods, we focus on the Geometric Entropy Minimization (GEM) algorithm [22], [23]. This algorithm estimates the ME set based on the k-nearest neighbor graph (k-NNG), which is shown to be the Uniformly Most Powerful Test at given level when the anomalies are drawn from an unknown mixture of known nominal density and uniform anomalous density [22]. A *key contribution* of this paper is the incorporation of the non-parametric GEM anomaly detection into a binary classifier under a non-parametric corrupt-data model.

The proposed framework, called the *Geometric-Entropy-Minimization regularized by Maximum Entropy Discrimination (GEM-MED)*, follows a *Bayesian* perspective. It is an extension of the well-established Maximum Entropy Discrimination (MED) approach proposed by Jaakkola et al. [2]. MED per-

forms Bayesian large margin classification via the maximum entropy principle and it subsumes SVM as a special case. The MED model can also solve the parametric anomaly detection [2] problem and has been extended to multitask classification [29]. A naive application of MED to robust classification might use a two-stage approach that implements an anomaly detector on the training set prior to training the MED classifier, which is sub-optimal. In this paper, we propose GEM-MED as a unified approach that jointly solves an anomaly detection and classification problem via the MED framework. The GEM-MED explicitly incorporates the anomaly detection false-alarm constraint and the mis-classification rate constraint into a maximum entropy learning framework. Unlike the two-stage approach, GEM-MED finds anomalies by investigating both the underlying sample distribution and the sample-label relationship, allowing anomalies in support vectors to be more effectively suppressed. As a Bayesian approach, GEM-MED requires no tuning parameter as compared to other anomaly-resistant classification approaches, such as ROD [13]. We demonstrate the superior performance of the GEM-MED anomaly-resistant classification approach over other robust learning methods on simulated data and on a real data set combining sensor failure. The real data set contains human-alone and human-leading-animal footsteps, collected in the field by an acoustic sensor array [30]–[32].

## B. Organization of the paper

What follows is a brief outline of the paper. In Section II, we review MED as a general framework to perform classification and other inference tasks. The proposed combined GEM-MED approach is presented in Section III. A variational implementation of GEM-MED is introduced in Section IV. Experimental results based on synthetic data and real data are presented in Section V. Our conclusions are discussed in Section VI.

## II. FROM MED TO GEM-MED: A GENERAL ROUTINE

Denote the training data set as  $\mathcal{D}_t := \{(y_n, \mathbf{x}_n)\}_{n \in T}$ , where each sample-pair  $(y_n, \mathbf{x}_n) \in \mathcal{Y} \times \mathcal{X} = \mathcal{D}$  are independent. Denote the feature set  $\mathcal{X} \subset \mathcal{R}^p$  and the label set as  $\mathcal{Y}$ . For simplicity, let  $\mathcal{Y} = \{-1, 1\}$ . The test data set is denoted as  $\mathcal{D}_s := \{\mathbf{x}_m\}_{m \in S}$ . We assume that  $\{(y_n, \mathbf{x}_n)\}_{n \in T}$  are i.i.d. realizations of random variable  $(Y, X)$  with distribution  $\mathcal{P}_t$ , conditional probability density  $p(X|Y = y, \Theta)$  and prior  $p(Y = y), y \in \mathcal{Y}$ , where  $\Theta$  is the set of unknown model parameters. We denote by  $p(Y = y, X; \Theta) = p(X|Y = y, \Theta)p(Y = y)$  the parameterized joint distribution of  $(Y, X)$ . The distribution of test data, denoted as  $\mathcal{P}_s$ , is defined similarly.  $\mathcal{P}_{nom}$  denotes the nominal distribution. Finally, we define the probability simplex  $\Delta_{\mathcal{Y} \times \mathcal{X}}$  over the space  $\mathcal{Y} \times \mathcal{X}$ .

### A. MED for classification and parametric anomaly detection

The Maximum entropy discrimination (MED) approach to learning a classifier was proposed by Jaakkola et al [2]. The MED approach is a Bayesian maximum entropy learning framework that can either perform conventional classification, when  $\mathcal{P}_t = \mathcal{P}_s = \mathcal{P}_{nom}$ , or anomaly detection, when  $\mathcal{P}_t \neq \mathcal{P}_s$ , and  $\mathcal{P}_t = \mathcal{P}_{nom}$ . In particular, assume that all parameters

in  $\Theta$  are random with prior distribution  $p_0(\Theta)$ . Then MED is formulated as finding the posterior distribution  $q(\Theta)$  that minimizes the relative entropy

$$\text{KL}(q(\Theta) \parallel p_0(\Theta)) := \int \log \left( \frac{q(\Theta)}{p_0(\Theta)} \right) q(d\Theta) \quad (1)$$

subject to a set of  $P$  constraints on the risk or loss:

$$\int \mathcal{L}_i(p, (y_n, \mathbf{x}_n); \Theta) q(d\Theta) \leq 0, \forall n \in T, 1 \leq i \leq P. \quad (2)$$

The constraint functions  $\{\mathcal{L}_i\}_{i=1}^P$  can correspond to losses associated with different type of errors, e.g. misdetection, false alarm or misclassification. For example, the classification task defines a parametric discriminant function  $\mathcal{F}_C : \Delta_{\mathcal{Y} \times \mathcal{X}} \times \mathcal{D} \rightarrow \mathbb{R}_+$  as

$$\mathcal{F}_C(p, (y_n, \mathbf{x}_n); \Theta) := \log p(Y = y_n | \mathbf{x}_n; \Theta) / p(Y \neq y_n | \mathbf{x}_n; \Theta).$$

In the case of the SVM classification, the loss function is defined as

$$\mathcal{L}_i = \mathcal{L}_C(p, (y_n, \mathbf{x}_n); \Theta) := [\xi_n - \mathcal{F}_C(p, (y_n, \mathbf{x}_n); \Theta)]. \quad (3)$$

Other definitions of discriminant functions are also possible [2].

An example of an anomaly detection test function  $\mathcal{L}_i = \mathcal{L}_D : \Delta_{\mathcal{X}} \times \mathcal{X} \rightarrow \mathbb{R}$ , is

$$\mathcal{L}_D(p, \mathbf{x}_n; \Theta) := -[\log p(\mathbf{x}_n; \Theta) - \beta], \quad (4)$$

where  $p(\mathbf{x}_n; \Theta)$  is the marginal likelihood  $p(\mathbf{x}_n; \Theta) = \sum_{y_n \in \mathcal{Y}} p(X = \mathbf{x}_n | Y = y_n, \Theta) p(Y = y_n)$ . The constraint function (4) has the interpretation as local entropy of  $X$  in the neighborhood of  $X = \mathbf{x}_n$ . Minimization of the *average* constraint function yields the minimal entropy anomaly detector [22], [23]. The solution to the minimization (2) yields a posterior distribution  $p(Y = y | \mathbf{x}_n, \bar{\Theta})$  where  $\bar{\Theta} := \Theta \cup \{\xi_n\} \cup \{\beta\}$ . This lead to a discrimination rule

$$y^* = \operatorname{argmin}_y \left\{ - \int \log p(y, \mathbf{x}_m; \bar{\Theta}) q(d\bar{\Theta}) \right\}, \mathbf{x}_m \in \mathcal{D}_s. \quad (5)$$

when applied to the test data  $\mathcal{D}_s$ .

The decision region  $\{\mathbf{x} \in \mathcal{X} | Y = y\}$  of MED can have various interpretations depending on the form of the constraint function (3) and (4). For the anomaly detection constraint (4), it is easily seen that the decision region is a  $\beta$ -level-set region for the marginal  $p(\mathbf{x}; \bar{\Theta})$ , denoted as  $\Psi_\beta := \{\mathbf{x}_n \in \mathcal{X} | \log p(\mathbf{x}_n; \bar{\Theta}) \geq \beta\}$ . Here  $\Psi_\beta$  is the *rejection region* associated with the test: declare  $\mathbf{x}_m \in \mathcal{D}_s$  as anomalous whenever  $\mathbf{x}_m \notin \Psi_\beta$ ; and declare it as nominal if  $\mathbf{x}_m \in \Psi_\beta$ . With a properly-constructed decision region, the MED model, as a projection of prior distribution  $p_0(\Theta)$  into this region, can provide performance guarantees in terms of the error rate or the false alarm rate and can result in improved accuracy [29], [33].

Similar to the SVM, the MED model readily handles non-parametric classifiers. For example, the discriminant function

$\mathcal{F}_C(p, (y, \mathbf{x}); \Theta)$  can take the form  $y[\Theta(\mathbf{x})]$  where  $\Theta = f$  is a random function, and  $f : \mathcal{X} \rightarrow \mathcal{Y}$  can be specified by a Gaussian process with Gaussian covariance kernel  $K(\cdot, \cdot)$ . More specifically,  $f \in \mathcal{H}$ , where  $\mathcal{H}$  is a Reproducing Kernel Hilbert Space (RKHS) associated with kernel function  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . See [29] for more detailed discussion.

MED utilizes a weighted ensemble strategy that can improve the classifier stability [2]. However, like SVM, MED is sensitive to anomalies in the training set.

### B. Robustified MED when there is an anomaly detection oracle

Assume an *oracle* exists that identifies anomalies in the training set. Using this oracle, partition the training set as  $\mathcal{D}_t = \mathcal{D}_t^{nom} \cup \mathcal{D}_t^{anm}$ , where  $(\mathbf{x}_n, y_n) \sim \mathcal{P}_{nom}$  if  $(\mathbf{x}_n, y_n) \in \mathcal{D}_t^{nom}$  and  $(\mathbf{x}_n, y_n) \not\sim \mathcal{P}_{nom}$ , if  $(\mathbf{x}_n, y_n) \in \mathcal{D}_t^{anm}$ . Given the oracle, one can achieve robust classification simply by constructing a classifier and an anomaly detector simultaneously on  $\mathcal{D}_t^{nom}$ . This results in a naive implementation of robustified MED as

$$\min_{q(\bar{\Theta}) \in \Delta_{\bar{\Theta}}} \text{KL}(q(\bar{\Theta}) \parallel p_0(\bar{\Theta})) \quad (6)$$

$$\text{s.t.} \int \mathcal{L}_C(p, (y_n, \mathbf{x}_n); \bar{\Theta}) q(d\bar{\Theta}) \leq 0, (\mathbf{x}_n, y_n) \in \mathcal{D}_t^{nom}, \quad (7)$$

$$\int \mathcal{L}_D(p, \mathbf{x}_n; \bar{\Theta}) q(d\bar{\Theta}) \leq 0, (\mathbf{x}_n, y_n) \in \mathcal{D}_t^{nom}, \quad (8)$$

where  $\bar{\Theta} = \Theta \cup \{\beta\} \cup \{\xi_n\}_{n \in T}$ , the large-margin error function  $\mathcal{L}_C$  is defined in (3) and the test function  $\mathcal{L}_D$  is defined in (4). The prior is defined as  $p_0(\bar{\Theta}) = p_0(\Theta) p_0(\beta) \prod_{n \in T} p_0(\xi_n)$ .

Of course, the oracle partition  $\mathcal{D}_t = \mathcal{D}_t^{nom} \cup \mathcal{D}_t^{anm}$  is not available *a priori*. The parametric estimator  $\hat{\Psi}_\beta$  of  $\Psi_\beta$  can be introduced in place of  $\mathcal{D}_t^{nom}$  in (6). However, the estimator  $\hat{\Psi}_\beta$  is difficult to implement and can be severely biased if there is model mismatch. Below, we propose an alternative nonparametric estimate of the decision region  $\Psi_\beta$  that learns the oracle partition.

## III. THE GEM-MED: MODEL FORMULATION

### A. Anomaly detection using minimal-entropy set

As an alternative to a parametric estimator of the level-set  $\Psi_\beta := \{\mathbf{x}_m \in \mathcal{X} | \log p(\mathbf{x}_m; \bar{\Theta}) \geq \beta\}$ , we propose to use a non-parametric estimator [34] based on the *minimal-entropy (ME) set*  $\Omega_{1-\beta}$ . The ME set  $\Omega_{1-\beta} := \arg \min_A \{H(A) | \int_A p(\mathbf{x}) d\mathbf{x} \geq \beta\}$  is referred as the *minimal-entropy-set of false alarm level*  $1 - \beta$ , where  $H(A) = -\int_A \log p(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$  is the Shannon entropy of the density  $p(\mathbf{x})$  over the region  $A$ . This minimal-entropy-set is equivalent to the *epigraph-set*  $\{A : \int_A p(\mathbf{x}) d\mathbf{x} \geq \beta\}$  as illustrated in Fig. 2.

Given  $\Omega_{1-\beta}$ , the ME anomaly test is as follows: a sample  $\mathbf{x}_n$  is declared anomalous if  $\mathbf{x}_n \notin \Omega_{1-\beta}$ ; and it is declared nominal, when  $\mathbf{x}_n \in \Omega_{1-\beta}$ . It is established in [22] that when  $p(\mathbf{x})$  is a known density, this test is a Uniformly Most Powerful

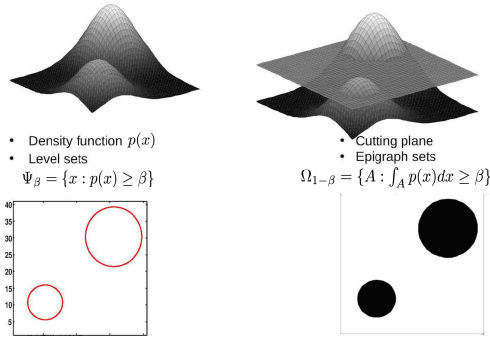


Fig. 2. The comparison of level-set (left) and the epigraph-set (right) w.r.t. two continuous density function  $p(x)$ . The minimum-entropy-set is computed based on the epigraph-set.

Test (UMPT) [35] at level  $\beta$  of the hypothesis  $H_0 : x \sim p(x)$  vs.  $H_1 : x \sim p(x) + \epsilon U(x)$ , where  $U(x)$  is the uniform density and  $\epsilon \in [0, 1]$  is an unknown mixture coefficient.

### B. The BP-kNNG implementation of GEM

Several methods have been proposed to empirically approximate the ME set  $\Omega_{1-\beta}$  including: kernel density estimation [21]; the  $k$ -point minimal spanning tree [36]; the leave-one-out  $k$ -nearest-neighbor graph [23]; and the average  $k$ -nearest-neighbor distance [37]. In [23], the bipartite  $k$ -nearest-neighbor (BP-kNN) based algorithm was proposed as an alternative approximation. The BP-kNN solves the following discrete optimization problem:

$$A_c^* \in \arg \min_{A_c \subset \mathcal{D}_t^{N,c}} L(A_c, \mathcal{D}_t^{M,c}),$$

$$\text{where } L(A_c, \mathcal{D}_t^{M,c}) := \sum_{\mathbf{x}_n \in A_c} d_k(\mathbf{x}_n, \mathcal{D}_t^{M,c}),$$

and where  $A_c$  is a set of distinct  $K = |T|(1 - \beta)$  points in  $\mathcal{D}_t^{N,c}$  (see Fig. 3 for illustration). It is shown in [23] that  $A_c^* = \hat{\Omega}_{1-\beta}$  is an asymptotically consistent estimator of the ME set. Equivalently, let  $\eta_n \in \{0, 1\}$  be the indicator function of the event  $\mathbf{x}_n \in A_c$  and define  $d_n := d_k(\mathbf{x}_n, \mathcal{D}_t^{M,c})$ . Then it can easily be shown that the algorithm in [23] finds the optimal binary variables  $\{\eta_n \in \{0, 1\} | \mathbf{x}_n \in \mathcal{D}_t^{N,c}\}$ ,  $n = 1, \dots, N$ , that minimize

$$\sum_{\mathbf{x}_n \in \mathcal{D}_t^{N,c}} \eta_n d_n \quad \text{subject to} \quad \sum_{\mathbf{x}_n \in \mathcal{D}_t^{N,c}} \eta_n \geq K. \quad (9)$$

This representation makes the BP-kNN implementation of GEM naturally adaptable to our GEM-MED framework. Specifically, the binary weights  $\eta_n \in \{0, 1\}$  are relaxed to continuous weights in the unit interval  $[0, 1]$  for all  $n \in T$ . After relaxation, the constraint in (9) becomes  $\sum_n \eta_n / |T| \geq \hat{\beta}$ , where  $\hat{\beta} = K / |T| = (1 - \beta) > 0$  is set so that the optimal solution  $\{\eta_n | \mathbf{x}_n \in A_c^*\}$  is feasible and the all-zero solution is infeasible. With the set of weights  $\{\eta_n\}_{n \in T}$ , the GEM problem in (9) can be transformed into a set of nonparametric constraints that fit the framework (6). This is discussed below.

### C. The GEM-MED as non-parametric robustified MED

Now we can implement the framework in (6). Denote  $\bar{\Theta} := \Theta \cup \{\hat{\beta}\} \cup \{\xi_n\}_{n \in T} \cup \{\eta_n\}_{n \in T} \cup \{\gamma_z\}_{z \in \{\pm 1\}}$ , where

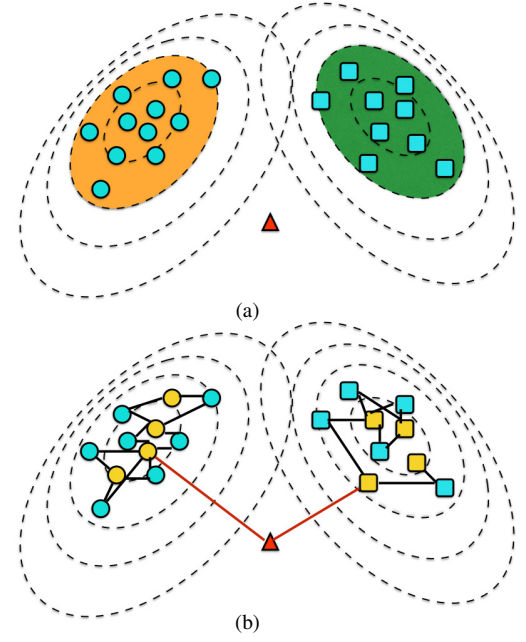


Fig. 3. Figure (a) illustrates ellipsoidal minimum entropy (ME) sets for two dimensional Gaussian features in the training set for class 1 (orange region) and class 2 (green region). These ME sets have coverage probabilities  $1 - \beta$  under each class distribution and correspond to the regions of maximal concentration of the densities. The blue disks and blue squares inside these regions correspond to the nominal training samples under class 1 and class 2, respectively. An outlier (in red triangle) falls outside of both of these regions. Figure (b) illustrates the bipartite 2-NN graph approach to identify the anomalous point, where the yellow disks and squares are reference samples in each class that are randomly selected from the training set. Note that the average 2-NN distance for anomalies should be significantly larger than that for the nominal samples.

$\Theta, \{\xi_n\}_{n \in T}$  are parameters as defined in (6),  $\{\eta_n\}_{n \in T}$  are weights in Sec. III-B and  $\hat{\beta}, \{\gamma_z\}_{z \in \{\pm 1\}}$  are variables to be defined later.

According to the objective function in (9), we specify the test function  $\tilde{\mathcal{L}}_D$  as

$$\begin{aligned} \tilde{\mathcal{L}}_D(\bar{\Theta}, \mathbf{y}; z, \mathbf{d}) &:= \tilde{\mathcal{L}}_D(\{\eta_n\}, \{\gamma_z\}, \mathbf{y}; z, \mathbf{d}) \\ &= \left( \sum_n \mathbb{1}\{y_n = z\} \eta_n d_n / |T| - \gamma_z \right), \quad z \in \{\pm 1\}, \end{aligned}$$

where  $\gamma_z \geq 0, z \in \{\pm 1\}$  is the threshold associated with  $d_n$  on  $\mathcal{D}_t \cap \{\mathbf{x}_n | y_n = z\}$ . Compared with (9), if  $\gamma_z = L_z^* + \epsilon$ , where  $L_z^*$  is the optimal value in (9) and  $\epsilon > 0$  is small enough, then for  $\{\eta_n\}_{n \in T}$  satisfying  $\tilde{\mathcal{L}}_D \leq 0$ , the region  $\{\mathbf{x}_n : \eta_n > \frac{1}{2}\}$  is concentrated on  $\hat{\Omega}_{1-\beta} \cap \{\mathbf{x}_n | y_n = z\}, z \in \{\pm 1\}$ .

As discussed in III-B, the constraint in (9) becomes the inequality constraint  $\sum_{n|y_n=z} \eta_n / |T| \geq \hat{\beta}$ .

Assuming that  $\bar{\Theta}$  is random with unknown distribution  $q(\bar{\Theta})$ , the above expected constraints becomes

$$\int \tilde{\mathcal{L}}_D(\bar{\Theta}, \mathbf{y}; z, \mathbf{d}) q(d\bar{\Theta}) \leq 0, z \in \{\pm 1\}, \quad (10)$$

$$\int \left[ \sum_{n:y_n=z} \eta_n / |T| \right] q(d\bar{\Theta}) \geq \hat{\beta}, z \in \{\pm 1\}. \quad (11)$$

The constraint (10) is referred as *the entropy constraint* and constraint (11) is the *epigraph constraint*. As discussed above, the region  $\{\mathbf{x}_n | \eta_n > \frac{1}{2}\}$  for  $q(\bar{\Theta})$  satisfying (10) and (11) is

concentrated on  $\widehat{\Omega}_{1-\beta} \cap \{\mathbf{x}_n | y_n = z\}$  in each class  $z \in \{\pm 1\}$  on average. With  $\mathcal{L}_D$ , the test constraint in (6) is replaced by (10) and (11).

For the classification part in (6), given  $\eta_n$  associated with each sample, the error constraints in (6) is replaced by *reweighted* error constraints

$$\int [\eta_n \mathcal{L}_C(p, (y_n, \mathbf{x}_n); \bar{\Theta})] q(d\bar{\Theta}) \leq 0, \quad n \in T,$$

with  $\mathcal{L}_C$  defined as in (3). Note that these constraints are applied to the entire training set. Summarizing, we have the following:

**Definition** The *Geometric-Entropy-Minimization Maximum-Entropy-Discrimination (GEM-MED)* method solves

$$\begin{aligned} \min_{q(\bar{\Theta}) \in \Delta_{\bar{\Theta}}} \quad & \text{KL}(q(\bar{\Theta}) \| p_0(\bar{\Theta})) \\ \text{s.t.} \quad & \int [\eta_n \mathcal{L}_C(p, (y_n, \mathbf{x}_n); \bar{\Theta})] q(d\bar{\Theta}) \leq 0, \quad n \in T, \\ & \int \tilde{\mathcal{L}}_D(\bar{\Theta}, \mathbf{y}; z, \mathbf{d}) q(d\bar{\Theta}) \leq 0, \quad z \in \{\pm 1\}, \\ & \int \left[ \sum_{n: y_n = z} \eta_n / |T| \right] q(d\bar{\Theta}) \geq \hat{\beta}, \quad z \in \{\pm 1\} \end{aligned} \quad (12)$$

where  $\bar{\Theta}$ ,  $\mathcal{L}_C$  and  $\tilde{\mathcal{L}}_D$  are defined as before.

#### IV. IMPLEMENTATION

##### A. Projected stochastic gradient descent algorithm

Note that (12) is a convex optimization w.r.t. the unknown distribution  $q(\bar{\Theta})$  [2], [38]. Therefore, it can be solved using the Karush-Kuhn-Tucker (KKT) conditions, which will result in a unique solution. We make the following simplifying assumptions under which our a computational algorithm is derived to solve (12).

- 1) Assume that a kernelized SVM is used for the classifier discriminant  $\mathcal{F}_C$  function. Following [29], [39], we assume that the decision function  $f$  follows a Gaussian random process on  $\mathcal{X}$ , i.e., a positive definite covariance kernel  $K(\mathbf{x}_i, \mathbf{x}_j)$  is defined for all  $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$  and all finite dimensional distributions, i.e., distributions of samples  $(f(\mathbf{x}_i))_{i \in T}$ , follow the multivariate normal distribution

$$(f(\mathbf{x}_i))_{i \in T} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad (13)$$

where  $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j \in T}$  is a specified covariance matrix. For example,  $K(\mathbf{x}_i, \mathbf{x}_j) := \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$  for Gaussian RBF kernel covariance function.

- 2) Assume a separable prior, as commonly used in Bayesian inference [2], [39], [40]

$$p_0(\bar{\Theta}) = p_0(\Theta) \prod_{n \in T} p_0(\xi_n) \prod_{n \in T} p_0(\eta_n) \prod_{z \in \{\pm 1\}} p_0(\gamma_z). \quad (14)$$

- 3) Assume that the hyperparameters  $\{\xi_n\}$  are exponential random variables and the indicator variables  $\{\eta_n\}$  are independent Bernoulli random variables,

$$p_0(\xi_n) \propto \exp(-c_\xi(1 - \xi_n)), \quad \xi_n \in (-\infty, 1], \quad n \in T;$$

$$p_0(\eta_n) = \text{Ber}(p_\eta)$$

$$\begin{aligned} \text{with } p_\eta &= \frac{1}{1 + \exp(-(a_\eta - \eta_n))} \\ &:= \sigma(a_\eta - \eta_n), \quad \eta_n \in \{0, 1\}, \quad n \in T; \\ p_0(\gamma_z) &= \delta_{\hat{\gamma}_z}(\gamma_z); \quad z \in \{\pm 1\}, \end{aligned} \quad (15)$$

where  $(a_\eta, c_\xi)$  are parameters and  $\hat{\gamma}_z$  is the upper bound estimate for minimal-entropy in each class  $z = \pm 1$  given by GEM algorithm.  $\sigma(x) = 1/(1 + \exp(-x))$  is the sigmoid function.

Now by solving the primal version of optimization problem (12), we have

**Theorem 4.1:** The GEM-MED problem in (12) is convex with respect to the unknown distribution  $q(\bar{\Theta})$  and the unique optimal solution is a *generalized Gibbs distribution* with the density:

$$q(d\bar{\Theta}) = \frac{1}{Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})} p_0(d\bar{\Theta}) \exp(-E(\bar{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})), \quad (16)$$

where

$$\begin{aligned} E(\bar{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) &:= E(\Theta, \hat{\beta}, \{\xi_n\}, \{\eta_n\}, \{\gamma_z\}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) \\ &= \sum_{n \in T} \lambda_n \eta_n \mathcal{L}_{C, \Theta, \xi_n} - \sum_{z \in \{\pm 1\}} \mu_z \tilde{\mathcal{L}}_{D, z} \\ &\quad - \sum_{z \in \{\pm 1\}} \kappa_z \sum_{n: y_n = z} \eta_n / |T| + \sum_{z \in \{\pm 1\}} \kappa_z \hat{\beta} \end{aligned}$$

with  $\bar{\Theta} = \Theta \cup \{\hat{\beta}\} \cup \{\xi_n\}_{n \in T} \cup \{\eta_n\}_{n \in T} \cup \{\gamma_{+1}, \gamma_{-1}\}$  and where the dual variables  $\boldsymbol{\lambda} = \{\lambda_n, n \in T\}$ ,  $\boldsymbol{\mu} = (\mu_z, z \in \pm 1)$  and  $\boldsymbol{\kappa} = (\kappa_z, z \in \pm 1)$  are all nonnegative.  $Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$  is the *partition function*, which is given as

$$Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \int \exp(-E(\bar{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})) p_0(d\bar{\Theta}). \quad (17)$$

The factor  $\mathcal{L}_{C, \Theta, \xi_n} := \mathcal{L}_C(\cdot; \Theta, \xi_n)$  is defined as in (3),  $\tilde{\mathcal{L}}_{D, z} := \tilde{\mathcal{L}}_D(\cdot; z, \cdot)$  is defined as in (10). See the Appendix Sec. A for a detailed derivation.

Moreover, we specify the error function as

$$\mathcal{L}_C(p, (y_n, \mathbf{x}_n); \Theta, \xi_n) := \xi_n - y_n f(\mathbf{x}_n), \quad (18)$$

where  $\Theta := f: \mathcal{X} \rightarrow \mathcal{Y}$  is a decision function associated with a nonparametric classifier as defined in Sec II-A.

Since the optimization problem is convex, we can equivalently solve a dual version of the optimization problem (12). In fact, we have the following result:

**Theorem 4.2:** Assume that (13), (14), (15) hold, the dual optimization problem is given as

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa} \geq 0} \quad & -\log Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}) \\ &= -\log \int \exp(-E(\bar{\Theta}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})) p_0(d\bar{\Theta}) \\ &= \sum_{n \in T} (\lambda_n + \log(1 - \lambda_n/c)) - \sum_{z \in \{\pm 1\}} \mu_z \hat{\gamma}_z + \hat{\beta} \sum_{z \in \{\pm 1\}} \kappa_z \\ &\quad - \log \int \exp\left(\frac{1}{2} Q(\mathbf{K} \odot (\mathbf{y}\mathbf{y}^T), (\boldsymbol{\lambda} \odot \boldsymbol{\eta}))\right) \end{aligned} \quad (19)$$

---

**Algorithm 1** The (kernel) GEM-MED algorithm

---

**Input:** The training set  $\mathcal{D}_t \subset \mathcal{X} \times \{\pm 1\}$  and the test set  $\mathcal{D}_s$ . The projection gradient step parameter  $\varphi, \psi, \tau > 0$ . Prior distribution and assumptions given as (13)-(15). The kernel function  $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is specified.

- 1: **Initialize:** Set  $\boldsymbol{\mu}_0 = \mathbf{0}, \boldsymbol{\kappa}_0 = \mathbf{0}$ .  $\lambda_0$  is set by applying conventional MED on  $\mathcal{D}$
- 2: **for**  $t = 1, \dots, T$  or until converge **do**
- 3:   Compute the gradient of log-partition function w.r.t  $\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t$  and  $\boldsymbol{\kappa}_t$ , respectively, i.e.  $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \lambda_n}$ ,  $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \mu_z}$  and  $\frac{\partial -\log Z(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t)}{\partial \kappa_z}$  according to the formula (23)-(25) where the expectation is approximated via Gibbs sampling described as above.
- 4:   Update  $\lambda_n, \mu_z$  and  $\kappa_z$  via projected gradient descent, i.e.

$$\begin{aligned} \lambda_{n,(t+1)} &= \text{proj}_{\{\lambda: 0 \leq \lambda \leq C_1\}} \left\{ \lambda_{n,t} - \varphi \frac{\partial \log Z(\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t)}{\partial \lambda_n} \right\} \\ n &\in T, \\ \mu_{z,(t+1)} &= \text{proj}_{\{\mu: \mu \geq 0\}} \left\{ \mu_{z,t} - \psi \frac{\partial \log Z(\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t)}{\partial \mu_z} \right\} \\ z &\in \{-1, +1\}, \\ \kappa_{z,(t+1)} &= \text{proj}_{\{\kappa: \kappa \geq 0\}} \left\{ \kappa_{z,t} - \tau \frac{\partial \log Z(\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t, \boldsymbol{\kappa}_t)}{\partial \kappa_z} \right\} \\ z &\in \{-1, +1\}, \end{aligned}$$

where  $\text{proj}_{\{x: 0 \leq x \leq C\}}\{w\} \equiv \min(\max(w, 0), C)$  defines the projection of  $x$  on the feasible set  $\{z : 0 \leq z \leq C\}$ .

5: **end for**

**Output:** Assign label for test sample  $\mathbf{x}_m \in \mathcal{D}_s$  as

$$y^* = \text{sign} \left\{ \sum_{n \in T} \hat{\eta}_n \lambda_n^* y_n K(\mathbf{x}_m, \mathbf{x}_n) \right\}, \quad \mathbf{x}_m \in \mathcal{D}_s$$

where  $\hat{\eta}_n = \mathbb{E}[\eta_n | f]$  at the final iteration of step 4.

---

Fig. 4. The proposed GEM-MED algorithm based on the projected stochastic gradient descent [41]. The gradient with respect to dual variables (20)-(23) can be approximated via Gibbs sampling as discussed in Sec. IV. The constraints on the dual variable  $\lambda_n \in [0, C_1]$  are imposed by a *clipping* procedure  $\text{proj}_{\{w: 0 \leq w \leq C\}}\{w\} = \min(\max(w, 0), C)$  that is applied on each Gibbs move, similarly to the C-SVM algorithm [42]. The parameters  $(\psi, \varphi, \tau)$  control the stepsize of the gradient descent algorithm.

$$\times p_0(\boldsymbol{\eta}) \exp(\boldsymbol{\eta}^T (-\boldsymbol{\mu} \otimes \mathbf{d} + \boldsymbol{\kappa} \otimes \mathbf{e})) d\boldsymbol{\eta} \quad (20)$$

where  $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$  are nonnegative dual variables as defined in (16),  $\mathbf{e}$  is the all 1's vector,  $\odot$  is Hadamard product,  $\otimes$  is the Kronecker product, respectively, and

$$Q(\mathbf{K}, \mathbf{x}) = \mathbf{x}^T \mathbf{K} \mathbf{x}$$

is the quadratic form associated with the kernel  $K$ .

See Appendix Sec. B for derivations of this result.

It is seen from (19) that the dual objective function is concave w.r.t. dual variables  $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$ . However, the integral in (20) is not closed form, so an explicit form as a quadratic optimization in SVM is not available. Nevertheless, the only coupling in (20) comes from the joint distribution  $q(f, \boldsymbol{\eta})$ . In particular, under the prior assumption (13), (14), (15), the optimal solution (16) satisfies

- 1)  $q(\bar{\Theta}) = q(f, \boldsymbol{\eta}) \prod_n q(\xi_n) q(\gamma_{+1}) q(\gamma_{-1})$  is factorized.
- 2)  $q(\boldsymbol{\eta} | f) = \prod_{n \in T} q(\eta_n | f)$ , i.e. the  $\{\eta_n, n \in T\}$  are conditional independent given the decision boundary function  $f$ . Moreover,

$$\begin{aligned} q(\eta_n | f) &= \text{Ber}(q_n), \\ \text{with } q_n &= \sigma(\rho_n \mathcal{F}_n(f)) \end{aligned} \quad (21)$$

where  $\rho_n := \log \frac{1-p_0(\eta_n=1)}{p_0(\eta_n=1)}$ ,  $\mathcal{F}_n(f) := \lambda_n [y_n f(\mathbf{x}_n) - 1] - \mu_{y_n} h_n + \kappa_{y_n} / |T|$ ,  $\sigma(\cdot)$  is the sigmoid function as (15).

3)  $f | \boldsymbol{\eta} \sim \mathcal{N}(f | \hat{f}_{\boldsymbol{\eta}, \boldsymbol{\lambda}}(\cdot), \mathbf{K})$ , where

$$\hat{f}_{\boldsymbol{\eta}, \boldsymbol{\lambda}}(\cdot) = \sum_{n \in T} \lambda_n \eta_n y_n K(\cdot, \mathbf{x}_n) \in \mathcal{H} \quad (22)$$

See Appendix Sec. C for details.

Given above results, we propose to use the *projected stochastic gradient descent* (PSGD, [41], [43]) algorithm to solve the dual optimization problem in (20). The gradient vectors of the dual objective function in (20) w.r.t.  $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa}$ , respectively, are computed as

$$\begin{aligned} \frac{\partial}{\partial \lambda_n} [-\log Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})] \\ = 1 - \mathbb{E}_{q(f, \boldsymbol{\eta})} [\eta_n y_n f(\mathbf{x}_n)] + \frac{c}{c - \lambda_n}, \quad n \in T; \end{aligned} \quad (23)$$

$$\begin{aligned} \frac{\partial}{\partial \mu_z} [-\log Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})] \\ = \mathbb{E}_{q(f, \boldsymbol{\eta})} \left\{ \sum_{n: y_n = z} \eta_n d_n \right\} - \hat{\gamma}_z, \quad z \in \{\pm 1\}; \end{aligned} \quad (24)$$



$$\begin{aligned} & \frac{\partial}{\partial \kappa_z} [-\log Z(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})] \\ &= \hat{\beta} - \frac{1}{|T|} \mathbb{E}_{q(f, \boldsymbol{\eta})} \left[ \sum_{n: y_n = z} \eta_n \right], \quad z \in \{\pm 1\}. \end{aligned} \quad (25)$$

Note that the expectation w.r.t.  $q(f, \boldsymbol{\eta})$  are approximated by Gibbs sampling with each conditional distribution given by (21), (22). For a detailed implementation of the Gibbs sampler, see the Appendix Sec. D.

A complete description of algorithm is presented in **Algorithm 1**. It is remarked that in (21) the probability of  $\{\eta_n = 0\}$  is proportional to the sum of margin of classification and negative local entropy value. The role of the dual variables  $(\eta_n, \mu_c)$  in (21) and (22) is to balance the classification margin  $y f(\cdot)$  and local entropy  $h$  in determining the anomalies.

### B. Prediction and detection on test samples

The GEM-MED classifier is similar to the standard MED classifier in (5):

$$\begin{aligned} y^* &= \operatorname{argmax}_y \left\{ \int y f(\mathbf{x}_m) q(f | \hat{\boldsymbol{\eta}}, \mathcal{D}_t) df \right\}, \\ &= \operatorname{sign} \left\{ \sum_{n \in T} \hat{\eta}_n \lambda_n^* y_n K(\mathbf{x}_m, \mathbf{x}_n) \right\} \quad \mathbf{x}_m \in \mathcal{D}_s. \end{aligned} \quad (26)$$

where  $\hat{\boldsymbol{\eta}}$  is the conditional mean estimator of  $\boldsymbol{\eta}$  given by Algorithm 1.

The GEM-MED was optimized on the training set to detect and mitigate anomaly corrupted training samples. When there are also anomalies in the test sample, an anomaly detection method can be applied independently to screen out these samples (at a given false positive rate) before applying GEM-MED to classify them. Such a two-stage approach to handling anomalies in the test sample is obviously not optimal. An optimal joint approach to handling anomalies in the training and test samples is worthwhile future direction which will not be investigated here.

## V. EXPERIMENTS

We illustrate the performance of the proposed GEM-MED algorithm on simulated data as well as on a real data collected in a field experiment. We compare the proposed GEM-MED with the SVM implemented by *LibSVM* [42] and the Robust-Outlier-Detection algorithm implemented with code obtained from the authors of [13]. For the simulated data experiment, a linear kernel SVM is implemented, and for the real data, a Gaussian RBF kernel SVM with kernel  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$  is implemented and the kernel parameter  $\gamma > 0$  is tuned via 5-fold-cross validation.

### A. Simulated experiment

For each class  $c \in \{\pm 1\}$ , we generate samples from the bivariate Gaussian distribution  $\mathcal{N}(\mathbf{m}_{+1}, \Sigma)$  and  $\mathcal{N}(\mathbf{m}_{-1}, \Sigma)$ , with mean  $\mathbf{m}_{-1} = (3, 3)$  and  $\mathbf{m}_{+1} = -\mathbf{m}_{-1}$  and common covariance  $\Sigma = \begin{bmatrix} 20 & 16 \\ 16 & 20 \end{bmatrix}$ . The sample follows the

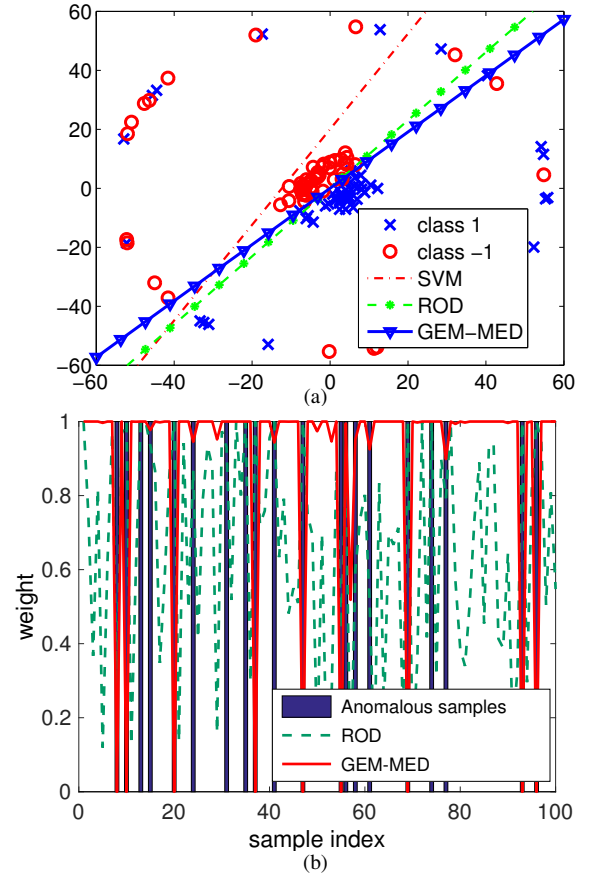


Fig. 5. (a) The classification decision boundary for SVM, ROD and GEM-MED on the simulated data set with two bivariate Gaussian distribution  $\mathcal{N}(\mathbf{m}_{+1}, \Sigma)$ ,  $\mathcal{N}(\mathbf{m}_{-1}, \Sigma)$  in the center and a set of anomalous samples for both classes distributed in a ring. Note that SVM is biased toward the anomalies (within outer ring support) and ROD and GEM-MED are insensitive to the anomalies. (b) The illustration of anomaly score  $\hat{\eta}_n$  for GEM-MED and ROD. The GEM-MED is more accurate than ROD in term of anomaly detection.

log-linear model  $\log p(y, \mathbf{x}; \bar{\Theta}) \propto 1/2 y(\mathbf{w}^T \mathbf{x} + b)$  where  $\bar{\Theta} = (\mathbf{w}, b)$ . A Gaussian prior was used as  $p_0(\bar{\Theta}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \sigma_w^2 \mathbf{I}) \mathcal{N}(b; 0, \sigma_b^2)$ .

We followed the same models as in [13]. In particular, the anomalies in the training set were drawn uniformly from a ring with an inner radius of  $R$  and outer radius  $R + 1$ , where  $R$  was assigned as one of the values  $[15, 35, 55, 75]$ . Define  $R$  to be the *noise level* of the data set, since the larger  $R$  the higher the discrepancy between the nominal distribution and the anomalous distribution. The samples then were labeled as  $\{0, 1\}$  with equal probability. The size of the training set was 100 for each class, and the ratio of anomaly samples was  $r_a$ . The test set contained 2000 uncorrupted samples from each class. See Fig. 5 (a) for a realization of the data set and the classifiers.

We first compare the classification accuracy of SVM, Robust-Outlier-Detection (ROD) with outlier parameter  $\rho$  and GEM-MED, under noise level  $R$  and a range of corruption rates  $r_a \in \{0.2, 0.3, 0.4, 0.5\}$ . We used the BP-kNNG implementation of GEM, where the k-nearest neighbor parameter  $k = 5$ . In the update of the GEM-MED dual variables  $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\kappa})$ , the learning rate  $(\varphi, \psi, \tau)$  is chosen based on a comparison of classification performance of the GEM-MED

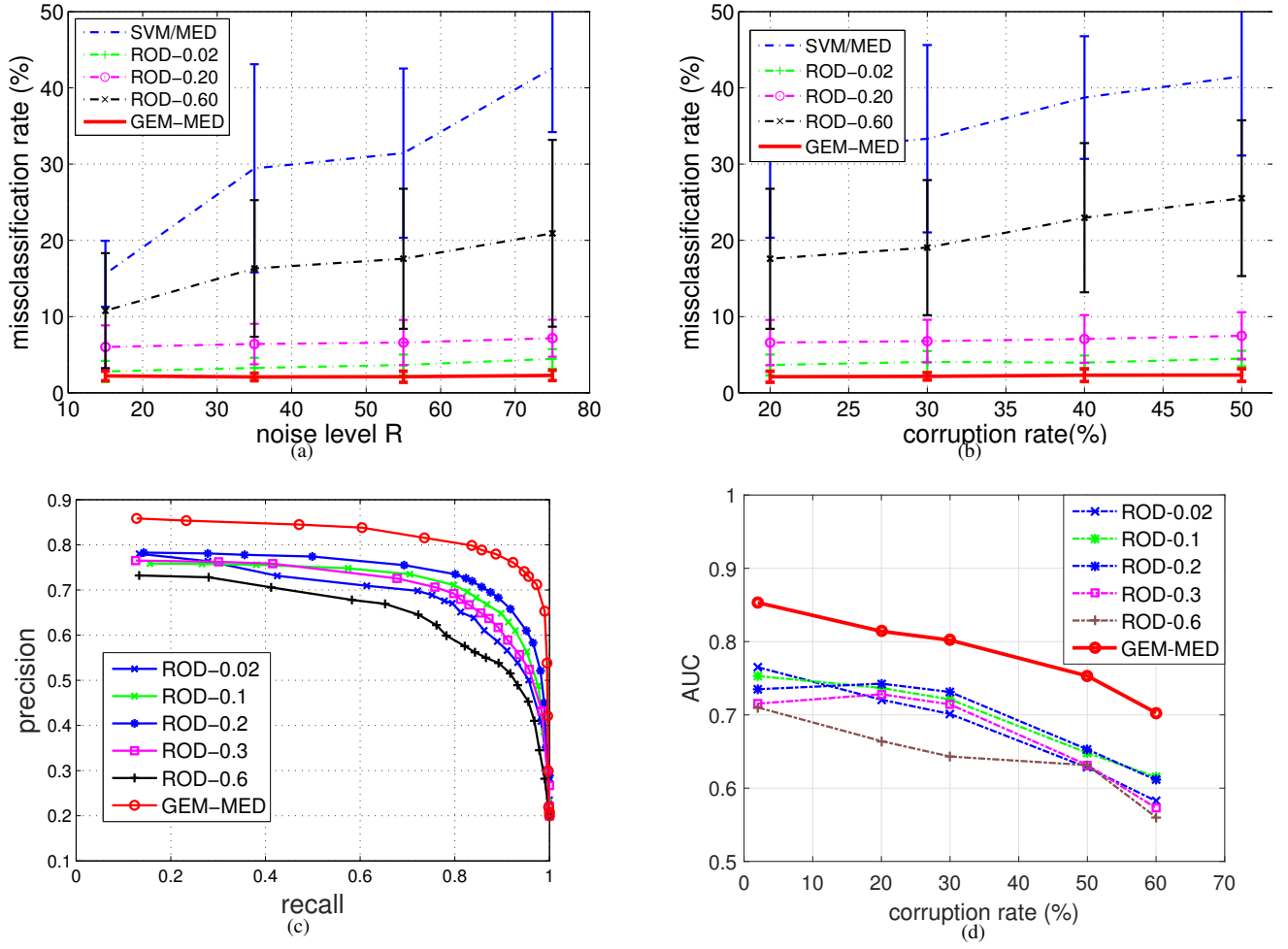


Fig. 6. (a) Miss-classification error (%) vs. noise level  $R$  for corruption rate  $r_a = 0.2$ . (b) Miss-classification error (%) vs. corruption rate  $\mathbb{E}[\eta]$  for ring-structured anomaly distribution having ring  $R = 55$ . (c) Recall-precision curve for GEM-MED and RODs on simulated data for corruption rate = 0.2. (d) The AUC vs. corruption rate  $r_a$  for GEM-MED and ROD with a range of outlier parameters  $\rho$ . From (a) and (b), GEM-MED outperforms both SVM/MED and ROD for various  $\rho$  in classification accuracy. From (c), under the same corruption rate, we see that GEM-MED outperforms ROD in terms of the precision-recall behavior. This due to the superiority of GEM constraints in enforcing anomaly penalties into the classifier. From (d), The GEM-MED outperforms RODs in terms of AUC for the range of investigated corruption rates.

under a range of noise levels  $R$  and corruption rates  $r_a$ , as shown in Fig. 7 (a)-(c). Note that when  $\varphi \in [1, 4] \times 10^{-3}$ ,  $\psi \in [1, 4] \times 10^{-2}$ ,  $\tau \in [1, 5] \times 10^{-2}$ , the performance of the GEM-MED is stable in terms of the averaged missclassification error and the variance. We fix  $(\varphi, \psi, \tau)$  in the stable range in the following experiments. For the ROD, we investigated a range of algorithm parameters, in particular outlier parameter  $\rho \in \{0.02, 0.2, 0.6\}$  for comparison, and we observed that the value  $\rho = 0.02$  gives the best classification performance regardless of the setting of  $R \in \{15, 35, 55, 75\}$  or  $r_a \in \{0.2, 0.3, 0.4, 0.5\}$ . Recall that the ROD parameter  $\rho$  is a fixed threshold that determines the proportion of anomalies, i.e., the proportion of nonzero  $\eta_n$  [13]. Compared to the ROD, the GEM-MED as a Bayesian method requires no tuning parameter to control the proportion of anomalies. In the experiments below, we compare the ROD for a range of outlier parameters  $\rho$  with GEM-MED for a single choice of  $(\varphi, \psi, \tau)$ , which were tuned via 5-fold-cross-validation of missclassification rate over 50 trial runs.

Fig. 6(a) shows the miss-classification error (%) versus

various noise level  $R$  (with  $r_a = 0.2$ ), and Fig. 6(b) shows the miss-classification error under different corruption rate settings (with  $R = 55$ ). In both experiments, GEM-MED outperforms ROD and SVM in terms of classification accuracy. Note that when the noise level or the corruption rate increases, the training data become less representative of the test data and the difference between their distributions increases. This causes a significant performance deterioration for the SVM/MED method, which is demonstrated in Fig. 6(a) and Fig. 6(b). Fig. 5 (b) also shows the bias of the SVM classifier towards the anomalies that lie in the ring. Comparing to GEM-MED and ROD in Fig 6(a) and Fig. 6(b), the former method is less sensitive to the anomalies. Moreover, since the GEM-MED model takes into account the marginal distribution for the training sample, it is more adaptive to anomalies in the training set, as compared to ROD, which does not use any prior knowledge about the nominal distribution but only relies on the predefined outlier parameter  $\rho$  to limit the training loss.

In Fig. 6(c) we compare the performance of GEM-MED and ROD in terms of precision vs recall for the same corruption



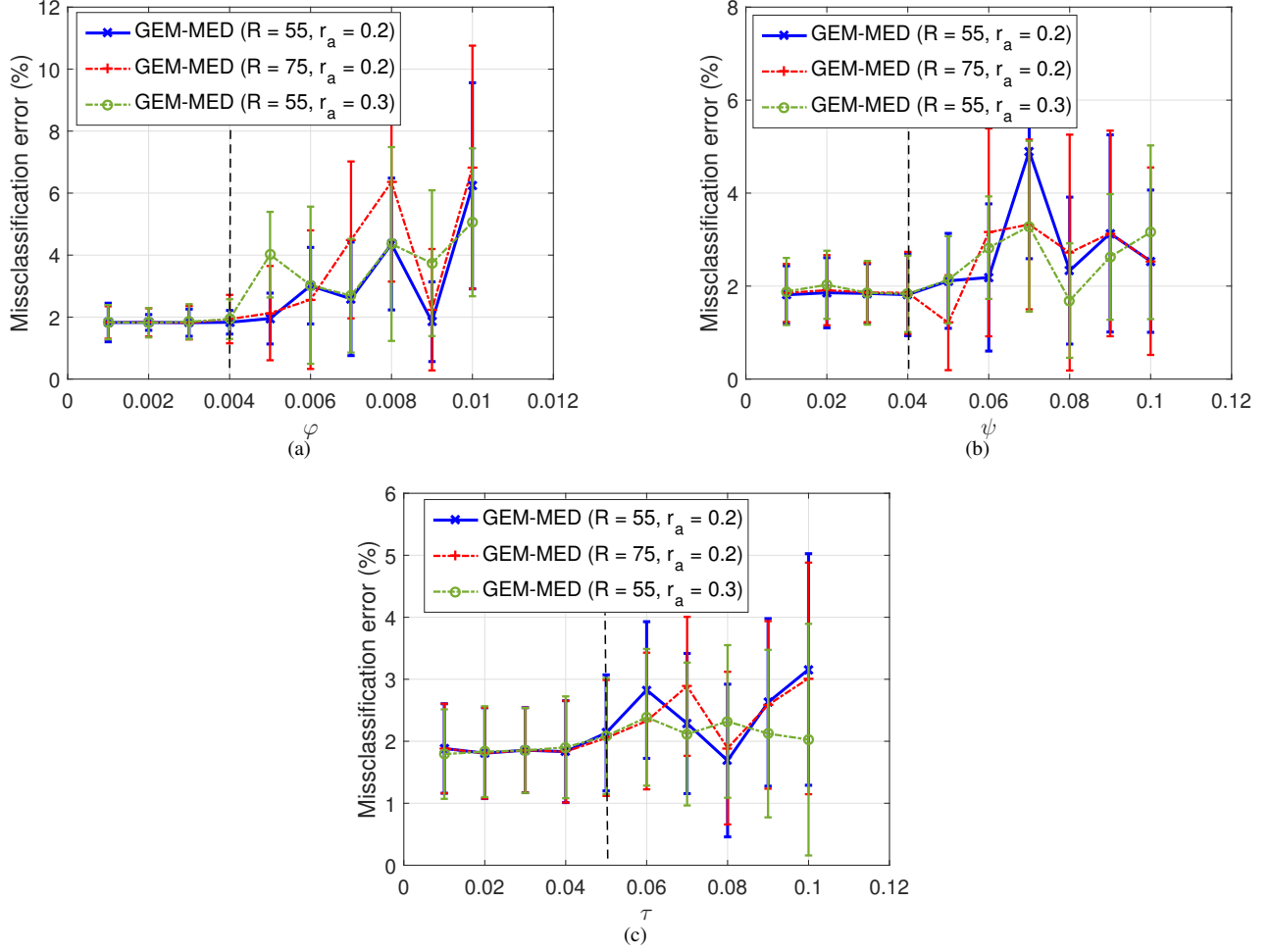


Fig. 7. The classification error of GEM-MED vs. (a) learning rates  $\varphi$ , when ( $\psi = 0.01, \tau = 0.02$ ); (b) vs.  $\psi$  when ( $\varphi = 0.001, \tau = 0.02$ ) and (c) vs.  $\tau$  when ( $\varphi = 0.001, \psi = 0.01$ ). The vertical dotted line in each plot separates the breakdown region (to the right) and the stable region of misclassification performance. These threshold values do not vary significantly as the noise level  $R$  and corruption rate  $r_a$  vary over the ranges investigated.

rate as in Fig. 6(a) and 6(b). In ROD and GEM-MED, the estimated weights  $\eta_n \in [0, 1]$  for each sample can be used to infer the likelihood of anomalies. In particular, in GEM-MED the corresponding latent variable estimate  $\hat{\eta}_n$  is obtained at the final iteration of the Gibbs sampling procedure, as described in Appendix Sec. D. Following the anomaly ranking procedure in [13], these anomaly scores are placed in ascending order. We compute the precision and recall using this ordering by averaging over 50 runs. Precision and recall are measures that are commonly used in data mining [44]:

$$\text{Precision} = \frac{|\{n : \eta_n \leq \rho_c\} \cap \{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|}{|\{n : \eta_n \leq \rho_c\}|}$$

$$\text{Recall} = \frac{|\{n : \eta_n \leq \rho_c\} \cap \{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|}{|\{n : (\mathbf{x}_n, y_n) \text{ are anomalous}\}|},$$

where the threshold  $\rho_c$  is a cut-off threshold that is swept over the interval  $[0, 1]$  to trace out the precision-recall curves in Fig. 6(c). It is evident from the figure that the proposed GEM-MED outlier resistant classifier has better precision-recall performance than ROD. Other corruption rates  $r_a$  lead to similar results. In Fig. 6(d), we compare the performance of GEM-MED, RODs under different corruption rates in terms of the Area Under the Curve (AUC), a commonly used

measure in data mining [44]. Similar to Fig. 6(c), the GEM-MED outperforms RODs in terms of AUC for the range of investigated corruption rates.

### B. Footstep classification experiment

The proposed GEM-MED method was evaluated on experiments on a real data set collected by the U.S. Army Research Laboratory [31], [32], [45]. This data set contains footstep signals recorded by a multisensor system, which includes four acoustic sensors and three seismic sensors. All the sensors are well-synchronized and operate in a natural environment, where the acoustic signal recordings are corrupted by environmental noise and intermittent sensor failures. The task is to discriminate between human-alone footsteps and human-leading-animal footsteps. We use the signals collected via four acoustic sensors (labeled sensor 1,2,3,4) to perform the classification. See Fig. 8. Note that the fourth acoustic sensor suffers from sensor failure, as evidenced by its very noisy signal record (bottom panel of Fig. 8). The data set involves 84 human-alone subjects and 66 human-leading-animal subjects. Each subject contains 24 75%-overlapping sample segments to capture temporal localized signal information. We randomly selected

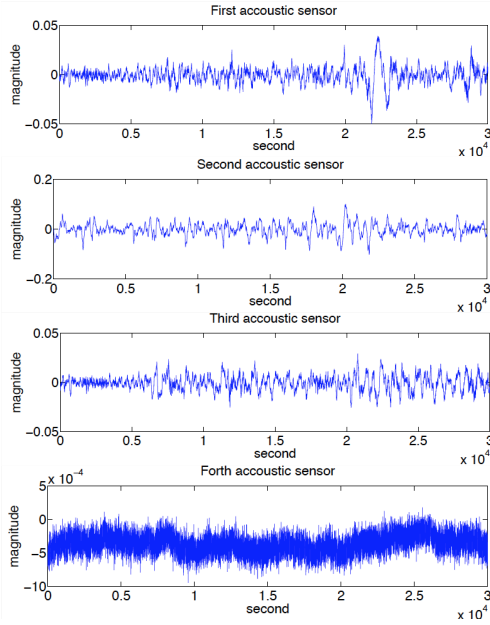


Fig. 8. A snapshot of human-alone footstep collected by four acoustic sensors.

25 subjects with 600 segments from each class as the training set. The test set contains the rest of the subjects. In particular, it contains 1416 segments from human-alone subjects and 984 segments from human-leading-animal subjects. A more detailed description of the dataset is given in [31], [32].

In a preprocessing step, for each segment, the time interval with strongest signal response is identified and signals within a fixed size of window (1.5 second) are extracted from the background. Fig. 9 shows the spectrogram (dB) of human-alone footsteps and human-leading-animal footsteps using the short-time Fourier transform [46], as a function of time (second) and frequency (Hz). The majority of the energy is concentrated in the low frequency band and the footstep periods differ between these two classes of signals. For features, we extract a mel-frequency cepstral coefficient (MFCC, [47]) vector using a 50 msec. window. Only the first 13 MFCC coefficients were retained, which were experimentally determined to capture an average 90% of the power in the associated cepstra. There are in total 150 windows for each segment, resulting in a matrix of MFCC coefficients of size  $13 \times 150$ . We reshaped the matrix of MFCC features to obtain a 1950 dimensional feature vector for each segment. We then apply PCA to reduce the dimensionality from 1950 to 50, while preserving 85% of the total power. The above procedures for preprocessing follows exactly from [45].

We compare the performance of kernel SVM, kernel MED, ROD for outlier parameter  $\rho \in [0.01, 1]$ , and GEM-MED by training on the four sensors individually as well as in combination. For the combined sensors we used an augmented feature vector of dimension 200 via feature concatenation. We used a Gaussian RBF kernel function for the matrix  $\mathbf{K}$  in the Gaussian process prior for the SVM decision function  $f$ . For the optimization of GEM-MED we used a separable prior and exponentially distributed hyperparameters, as indicated by (14) and (15). Finally, the BP-kNNG implementation of GEM

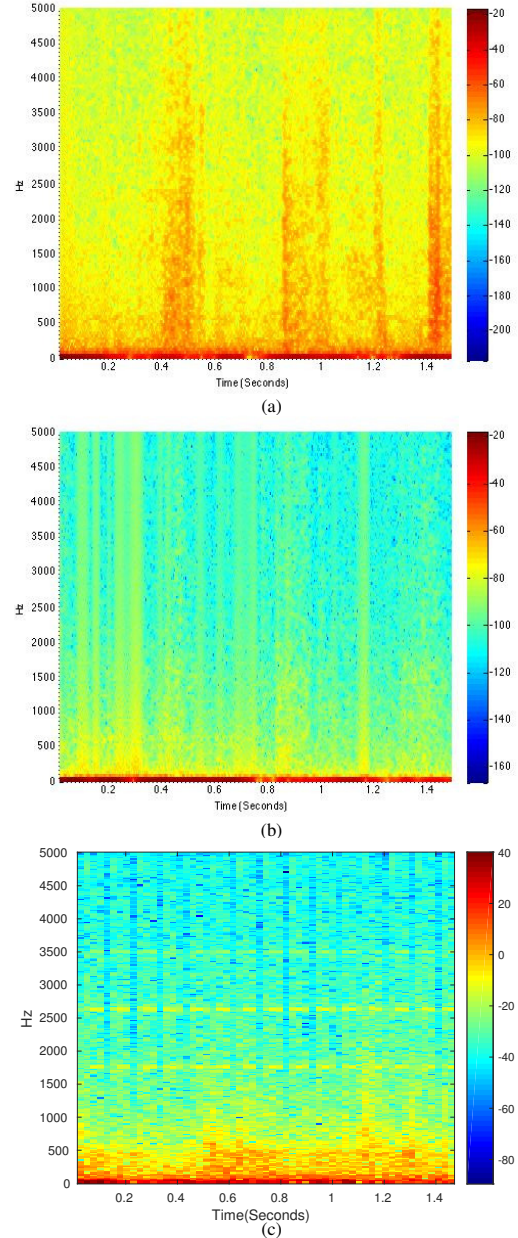


Fig. 9. The power spectrogram (dB) vs. time (sec.) and frequency (Hz.) for a human-alone footstep (a) and a human-leading-animal footstep (b). Observe that the period of periodic footstep is a discriminative feature that separates these two signals. (c) shows a corrupted human-alone footstep due to sensor malfunctioning.

was applied on the training samples in the MFCC feature space with  $k = 10$  nearest neighbors. The threshold  $\vartheta$  is set using the Leave-One-Out resampling strategy [22], where each holdout sample corresponds to an entire segment.

Note that all classifiers were learned from a corrupted training set. Since the test set is also corrupted we used an anomaly detection algorithm (GEM with 5% false alarm rate) to produce a test set with few anomalies, called the nominal test set. This allows us to report the performance of the various algorithms on both the clean test data and on the corrupted test data. Table II shows the classification accuracy of the methods (trained on the training set alone) applied to nominal test set and Table III shows the result on the entire corrupted test

TABLE II

CLASSIFICATION ACCURACY ON NOMINAL (CLEAN) TEST SET FOR FOOTSTEP EXPERIMENT WITH DIFFERENT SENSOR COMBINATIONS, WITH THE BEST PERFORMANCE SHOWN IN **BOLD**.

Classification Accuracy (%) mean $\pm$ standard error						
sensor no.	kernel SVM	kernel MED	ROD-0.02	ROD-0.2	GEM + SVM	GEM-MED
1	71.2 $\pm$ 8.2	71.1 $\pm$ 5.3	73.7 $\pm$ 3.7	76.0 $\pm$ 2.5	72.5 $\pm$ 4.2	<b>78.4 <math>\pm</math> 3.3</b>
2	60.8 $\pm$ 12.5	62.3 $\pm$ 10.2	71.5 $\pm$ 7.3	76.5 $\pm$ 5.3	70.3 $\pm$ 2.5	<b>82.1 <math>\pm</math> 3.1</b>
3	60.5 $\pm$ 14.2	60.0 $\pm$ 13.1	63.2 $\pm$ 5.4	<b>67.6 <math>\pm</math> 4.2</b>	56.5 $\pm$ 3.5	66.8 $\pm$ 4.5
4	59.6 $\pm$ 10.1	58.4 $\pm$ 8.2	71.8 $\pm$ 7.2	73.2 $\pm$ 4.2	76.5 $\pm$ 2.7	<b>80.1 <math>\pm</math> 3.1</b>
1,2,3,4	75.9 $\pm$ 7.5	78.6 $\pm$ 5.1	79.2 $\pm$ 3.7	79.8 $\pm$ 2.5	75.2 $\pm$ 3.3	<b>84.0 <math>\pm</math> 2.3</b>

TABLE III

CLASSIFICATION ACCURACY ON THE ENTIRE (CORRUPTED) TEST SET FOR FOOTSTEP EXPERIMENT WITH DIFFERENT SENSOR COMBINATIONS, WITH THE BEST PERFORMANCE SHOWN IN **BOLD**.

Classification Accuracy (%) mean $\pm$ standard error						
sensor no.	kernel SVM	kernel MED	ROD-0.02	ROD-0.2	GEM + SVM	GEM-MED
1	65.2 $\pm$ 10.6	65.8 $\pm$ 10.2	68.5 $\pm$ 8.3	70.0 $\pm$ 6.8	70.2 $\pm$ 5.5	<b>72.5 <math>\pm</math> 4.8</b>
2	54.9 $\pm$ 11.8	55.2 $\pm$ 11.0	63.2 $\pm$ 9.8	68.1 $\pm$ 7.5	68.5 $\pm$ 7.8	<b>76.3 <math>\pm</math> 3.9</b>
3	50.7 $\pm$ 10.0	52.0 $\pm$ 10.5	56.8 $\pm$ 8.5	56.9 $\pm$ 7.3	56.5 $\pm$ 3.5	<b>60.1 <math>\pm</math> 5.3</b>
4	57.0 $\pm$ 12.3	57.5 $\pm$ 12.1	69.6 $\pm$ 9.2	69.8 $\pm$ 5.1	70.2 $\pm$ 4.2	<b>75.0 <math>\pm</math> 4.0</b>
1,2,3,4	70.8 $\pm$ 8.8	71.0 $\pm$ 8.5	73.6 $\pm$ 7.2	74.8 $\pm$ 6.9	75.1 $\pm$ 3.3	<b>76.8 <math>\pm</math> 2.5</b>

set. For ROD only  $\rho = 0.02$  and  $\rho = 0.20$  are shown; it was determined that  $\rho = 0.20$  achieves the best performance in the range  $\rho \in [0.01, 1]$ . In Table II, it is seen that the GEM-MED method outperforms the ROD- $\rho$  algorithms for all values of  $\rho$  as a function of classification accuracy when individual sensors 1,2,4 are used. Notice that when used alone neither kernel MED nor kernel SVM is resistant to the sensor failures in the training set, which explains their poor accuracy in sensor 3 and sensor 4. Also in the column *GEM+MED* of Table II, we first trained a GEM anomaly detector to screen out 5% of the noisy training set, then trained a MED classifier on the rest of the training data. Note that GEM-MED learns both the detector and the classifier jointly on noisy training data. Table II shows that the two stage training approach has poor performance in highly corrupted sensors 3 and 4. This is due to the fact that when the GEM detector is learned without inferring the classification margin, it cannot effectively limit the negative influence of those corrupted samples that are close to the class boundary. In Table III, we show the classification accuracy when both the nominal and anomalous test samples are involved in evaluation. We observe a performance degradation for all methods due to the irregularity of the outliers in the test set. In spite of this, the GEM-MED maintains a superior performance over all other methods. This reflects the superiority of the proposed

TABLE IV

ANOMALY DETECTION ACCURACY WITH DIFFERENT SENSORS, WITH THE BEST PERFORMANCE SHOWN IN **BOLD**.

Anomaly Detection Accuracy (%) mean $\pm$ standard error			
sensor no.	ROD-0.02	ROD-0.2	GEM-MED
1	30.2 $\pm$ 1.3	59.0 $\pm$ 3.5	<b>70.5 <math>\pm</math> 1.3</b>
2	23.5 $\pm$ 2.6	<b>63.5 <math>\pm</math> 2.8</b>	63.4 $\pm$ 2.5
3	5.3 $\pm$ 1.4	48.1 $\pm$ 3.3	<b>72.8 <math>\pm</math> 1.5</b>
4	22.8 $\pm$ 3.2	65.2 $\pm$ 4.2	<b>88.1 <math>\pm</math> 2.1</b>
1, 2, 3, 4	38.5 $\pm$ 6.3	63.3 $\pm$ 5.5	<b>88.5 <math>\pm</math> 4.1</b>

joint classification and detection approach of GEM-MED as compared with *GEM + MED* approach.

Table IV compares the anomaly detection accuracies on both *training and test data* for ROD and GEM-MED, where the accuracy is computed relative to ground truth anomalies. Note that GEM-MED has significant improvement in accuracy over ROD when trained individually on sensors 1,3,4, respectively, and when trained on all of the combined sensors. When trained on sensor 2 alone, the accuracies of GEM-MED and ROD-0.2

are essentially equivalent. In sensor 2 the anomalies appear to occur in concentrated bursts and we conjecture that a GEM-MED model that accounts for clustered and dependent anomalies may be able to do better. Such an extension is left to future work.

## VI. CONCLUSION

In this paper we proposed a unified GEM-MED approach for anomaly-resistant classification. We demonstrated its performance advantages in terms of both classification accuracy and detection rate on a simulated data set and on a real footstep data set, as compared to an anomaly-blind Ramp-Loss-based classification method (ROD). Further work could include generalization to the setting of multiple sensor types where anomalies exist in both training and test sets.

## VII. ACKNOWLEDGMENT

This work was supported in part by the U.S. Army Research Lab under ARO grant WAI1NF-11-1-103A1. We also thanks Xu LinLi and Kumar Sricharan for their inputs on this work.

## APPENDIX

### A. Derivation of theorem 4.1

*Proof:* The proof of the convexity of the problem can be seen in chapter 12 of the standard textbook [2], since the problem is with respect to the distribution  $q$ . The uniqueness of the solution follows directly from the fact that the problem is convex.

The Lagrangian function is given as

$$\begin{aligned} \mathcal{L}(q, \lambda, \mu, \nu) &= \mathbb{E}_q [\log q - \log p_0] + \sum_{n \in T} \lambda_n \mathbb{E}_q [\eta_n \mathcal{L}_C] - \sum_{z \in \{\pm 1\}} \mu_z \mathbb{E}_q [\tilde{\mathcal{L}}_{D,z}] \\ &\quad - \sum_{z \in \{\pm 1\}} \kappa_z \mathbb{E}_q \left[ \sum_{n: y_n = z} \eta_n / |T| - \hat{\beta} \right] \end{aligned}$$

with dual variables  $\lambda = \{\lambda_n, n \in T\} \succeq \mathbf{0}$ ,  $\mu = (\mu_z, z \in \pm 1) \succeq \mathbf{0}$  and  $\nu \geq 0$ .

Then the result follows directly from solving a system of equations according to the KKT condition. ■

### B. Derivation of theorem 4.2

*Proof:* According to [2], the dual optimization is given as

$$\begin{aligned} \max_{\lambda, \mu, \kappa \geq 0} & -\log Z(\lambda, \mu, \kappa) \\ &= -\log \prod_{n \in T} \int \exp(-c(1 - \xi_n) - \lambda_n \xi_n) d\xi_n \\ &\quad \times \int \exp\left(-\frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} + \sum_n \lambda_n \eta_n y_n f_n\right) d\mathbf{f} \\ &\quad \times p_0(\boldsymbol{\eta}) \exp\left(-\sum_{z \in \{\pm 1\}} \mu_z \sum_{n: z} \eta_n d_n + \sum_{z \in \{\pm 1\}} \mu_z \hat{\gamma}_z\right. \end{aligned}$$

$$\begin{aligned} &\quad \left. + \sum_{z \in \{\pm 1\}} \kappa_z \sum_{n: z} \eta_n + \sum_{z \in \{\pm 1\}} \kappa_z \hat{\beta}\right) d\boldsymbol{\eta} \\ &= \sum_{n \in T} (\lambda_n + \log(1 - \lambda_n/c)) - \sum_{z \in \{\pm 1\}} \mu_z \hat{\gamma}_z - \left(\sum_{z \in \{\pm 1\}} \kappa_z\right) \hat{\beta} \\ &\quad - \log \int \exp\left(\frac{1}{2} Q(\mathbf{K}, (\lambda \odot \boldsymbol{\eta} \odot \mathbf{y})) + \boldsymbol{\eta}^T (-\boldsymbol{\mu} \otimes \mathbf{d} + \boldsymbol{\kappa} \otimes \mathbf{e})\right) \\ &\quad \times p_0(\boldsymbol{\eta}) d\boldsymbol{\eta} \end{aligned}$$

where

$$\begin{aligned} Q(\mathbf{K}, \mathbf{x}) &= \mathbf{x}^T \mathbf{K} \mathbf{x} \\ Q(\mathbf{K}, (\lambda \odot \boldsymbol{\eta})) &:= (\lambda \odot \boldsymbol{\eta})^T \mathbf{K} (\lambda \odot \boldsymbol{\eta}) \\ &= \boldsymbol{\lambda}^T (\mathbf{K} \odot (\boldsymbol{\eta} \boldsymbol{\eta}^T)) \boldsymbol{\lambda} \\ &= Q(\mathbf{K}(\boldsymbol{\eta}), \boldsymbol{\lambda}). \end{aligned}$$

■

### C. Derivation of (21), (22)

*Proof:* The expression for  $q(\bar{\Theta})$  is given as

$$\begin{aligned} q(\bar{\Theta}) &\propto \exp\left(-\frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} + \sum_n \lambda_n \eta_n y_n f_n\right) \\ &\quad \times p_0(\boldsymbol{\eta}) \exp\left(-\sum_{z \in \{\pm 1\}} \mu_z \sum_{n: z} \eta_n d_n + \sum_{z \in \{\pm 1\}} \kappa_z \sum_{n: z} \eta_n\right) \\ &\quad \times \prod_{n \in T} \exp(-c + (c - \lambda_n) \xi_n) \\ &= q(\mathbf{f}, \boldsymbol{\eta}) \prod_n q(\xi_n) \end{aligned}$$

Given all  $\eta_n, n \in T$ ,

$$\begin{aligned} q(\mathbf{f}|\boldsymbol{\eta}) &\propto \exp\left(-\frac{1}{2} \mathbf{f}^T \mathbf{K}^{-1} \mathbf{f} + \sum_n (\lambda_n \eta_n) f_n\right) \\ &= \exp\left(-\frac{1}{2} (\mathbf{f} - \mathbf{K}(\boldsymbol{\lambda} \odot \boldsymbol{\eta} \odot \mathbf{y}))^T \mathbf{K}^{-1} (\mathbf{f} - \mathbf{K}(\boldsymbol{\lambda} \odot \boldsymbol{\eta} \odot \mathbf{y}))\right) \\ &= \mathcal{N}(\mathbf{K}(\boldsymbol{\lambda} \odot \boldsymbol{\eta} \odot \mathbf{y}), \mathbf{K}). \end{aligned}$$

On the other hand, given  $\mathbf{f}, \boldsymbol{\eta} = (\eta_n, n \in T)$  are fully separated in above formula, therefore  $q(\boldsymbol{\eta}|\mathbf{f}) = \prod_n q(\eta_n|\mathbf{f})$ . ■

### D. Implementation of Gibbs sampler

We implement a Gibbs sampler [48] to estimate  $\mathbb{E}_{q(\mathbf{f}, \boldsymbol{\eta})} [G(\mathbf{f}, \boldsymbol{\eta})]$ , where  $G$  is a general function of  $\mathbf{f}$  and  $\boldsymbol{\eta}$ , as expressed in (23), (24), (25). The following procedure is applied iteratively

- Initialization: Set  $\hat{\boldsymbol{\eta}}_0 = [1, \dots, 1]^T$  and set a fixed dual parameter  $(\lambda, \mu, \kappa)$ . Let  $G_0 = 0$ .
- For each  $t = 1, 2, \dots, T_G$  or until convergence
  - 1) Given  $\hat{\boldsymbol{\eta}}_{t-1} = (\hat{\eta}_{n,t-1})$ , generate decision value  $f_t(\mathbf{x}_n), n = 1, \dots, N$  according to the Gaussian process (??) with mean function  $\hat{f}_t(\cdot) = \sum_{n \in T} \lambda_n \hat{\eta}_{n,t-1} y_n K(\cdot, \mathbf{x}_n)$ .
  - 2) Given  $\{f_t(\mathbf{x}_n)\}_{1 \leq n \leq N}$ , for  $r = 1, \dots, N_r$ ,

a) generate latent variables  $\eta_{n,t}^{(r)} \in \{0, 1\}$  according to the Bernoulli distribution with parameter as in (21) for each  $n$  independently.

3) Compute the sample mean of  $\hat{\eta}_{n,t} = \frac{1}{N_r} \sum_{r=1}^{N_r} \eta_{n,t}^{(r)} \in [0, 1], n = 1, \dots, N$ . Let  $\hat{\boldsymbol{\eta}}_t = (\hat{\eta}_{n,t})_{1 \leq n \leq N}$ .

4) Evaluate  $G_t$  via  $G_t = \frac{t-1}{t} G_{t-1} + \frac{1}{t} G(f_t, \hat{\boldsymbol{\eta}}_t)$

- Output the approximate expectation  $\hat{\mathbb{E}}_{q(f, \boldsymbol{\eta})} [G(f, \boldsymbol{\eta})] = G_T$  as well as the mean estimate  $\hat{\boldsymbol{\eta}}_T$  and  $\hat{f}_T(\mathbf{x}_n), 1 \leq n \leq N$  when the Gibbs chain process becomes stationary.

## REFERENCES

- [1] B. Schölkopf and A. J. Smola, *Learning with kernels*. The MIT Press, 2002.
- [2] T. Jaakkola, M. Meila, and T. Jebara, “Maximum entropy discrimination,” *Advances in Neural Information Processing Systems*, 1999.
- [3] P. L. Bartlett and S. Mendelson, “Rademacher and Gaussian complexities: Risk bounds and structural results,” *The Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2003.
- [4] O. Bousquet and A. Elisseeff, “Stability and generalization,” *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.
- [5] R. Chen, J.-M. Park, and K. Bian, “Robust distributed spectrum sensing in cognitive radio networks,” in *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE, 2008.
- [6] G. Ding, J. Wang, Q. Wu, L. Zhang, Y. Zou, Y.-D. Yao, and Y. Chen, “Robust spectrum sensing with crowd sensors,” *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3129–3143, 2014.
- [7] M. Yang, L. Xu, M. White, D. Schuurmans, and Y.-I. Yu, “Relaxed clipping: A global training method for robust regression and classification,” *Advances in neural information processing systems*, pp. 2532–2540, 2010.
- [8] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.
- [9] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.
- [10] D. E. Tyler, “Robust statistics: Theory and methods,” *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 888–889, 2008.
- [11] Q. Song, W. Hu, and W. Xie, “Robust support vector machine with bullet hole image classification,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 32, no. 4, pp. 440–448, 2002.
- [12] N. Krause and Y. Singer, “Leveraging the margin more carefully,” *Proceedings of the twenty-first international conference on Machine learning*, p. 63, 2004.
- [13] L. Xu, K. Crammer, and D. Schuurmans, “Robust support vector machine training via convex outlier ablation,” *AAAI*, vol. 6, pp. 536–542, 2006.
- [14] Y. Wu and Y. Liu, “Robust truncated hinge loss support vector machines,” *Journal of the American Statistical Association*, vol. 102, no. 479, 2007.
- [15] L. Wang, H. Jia, and J. Li, “Training robust support vector machine with smooth ramp loss in the primal space,” *Neurocomputing*, vol. 71, no. 13, pp. 3020–3025, 2008.
- [16] H. Masnadi-Shirazi and N. Vasconcelos, “On the design of loss functions for classification: theory, robustness to outliers, and savageboost,” *Advances in neural information processing systems*, pp. 1049–1056, 2009.
- [17] P. M. Long and R. A. Servedio, “Random classification noise defeats all convex potential boosters,” *Machine Learning*, vol. 78, no. 3, pp. 287–304, 2010.
- [18] P. A. Forero, V. Kekatos, and G. B. Giannakis, “Robust clustering using outlier-sparsity regularization,” *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4163–4177, 2012.
- [19] G. Ding, Q. Wu, Y.-D. Yao, J. Wang, and Y. Chen, “Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions,” *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 126–136, 2013.
- [20] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, “Support vector method for novelty detection,” *Advances In Neural Information Processing Systems*, vol. 12, pp. 582–588, 1999.
- [21] C. D. Scott and R. D. Nowak, “Learning minimum volume sets,” *The Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.
- [22] A. O. Hero, “Geometric entropy minimization (GEM) for anomaly detection and localization,” *Advances in Neural Information Processing Systems*, pp. 585–592, 2006.
- [23] K. Sricharan and A. Hero, “Efficient anomaly detection using bipartite k-NN graphs,” *Advances in Neural Information Processing Systems*, pp. 478–486, 2011.
- [24] J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with structural correspondence learning,” *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128, 2006.
- [25] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 193–200.
- [26] S. J. Pan and Q. Yang, “A survey on transfer learning,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [27] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.
- [28] H. Daume III and D. Marcu, “Domain adaptation for statistical classifiers,” *Journal of Artificial Intelligence Research*, pp. 101–126, 2006.
- [29] T. Jebara, “Multitask sparsity via maximum entropy discrimination,” *The Journal of Machine Learning Research*, vol. 12, pp. 75–110, 2011.
- [30] T. Damarla, A. Mehmood, and J. Sabatier, “Detection of people and animals using non-imaging sensors,” *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.
- [31] T. Damarla, “Seismic and ultrasonic data analysis for characterizing people and animals,” *SPIE Defense, Security, and Sensing*, 2012.
- [32] P.-S. Huang, T. Damarla, and M. Hasegawa-Johnson, “Multi-sensory features for personnel detection at border crossings,” *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.
- [33] J. Zhu, N. Chen, and E. P. Xing, “Infinite latent SVM for classification and multi-task learning,” *Advances in Neural Information Processing Systems*, pp. 1620–1628, 2011.
- [34] L. Wasserman, *All of Nonparametric Statistics*. Springer, 2010.
- [35] L. L. Scharf, *Statistical signal processing*. Addison-Wesley Reading, MA, 1991, vol. 98.
- [36] A. O. Hero and O. J. Michel, “Asymptotic theory of greedy approximations to minimal k-point random graphs,” *Information Theory, IEEE Transactions on*, vol. 45, no. 6, pp. 1921–1938, 1999.
- [37] J. Root, J. Qian, and V. Saligrama, “Learning efficient anomaly detectors from k-NN graphs,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015, pp. 790–799.
- [38] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [39] J. Zhu, N. Chen, and E. P. Xing, “Bayesian inference with posterior regularization and applications to infinite latent SVMs,” *Journal of Machine Learning Research*, vol. 15, pp. 1799–1847, 2014.
- [40] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [41] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.
- [42] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [43] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [44] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.
- [45] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, “Robust multi-sensor classification via joint sparse representation,” *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pp. 1–8, 2011.
- [46] E. Sejdić, I. Djurović, and J. Jiang, “Time–frequency feature representation using energy concentration: An overview of recent advances,” *Digital Signal Processing*, vol. 19, no. 1, pp. 153–183, 2009.
- [47] P. Mermelstein, “Distance measures for speech recognition, psychological and instrumental,” *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.
- [48] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.